Mini-Review

Genome and proteome analysis of *Clostridium felsineum DSM 794*

Daria A. Grek[1*]

[1*] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia

ABSTRACT
Motivation:
Results:
Availability:
Contact: dgrek@fbb.msu.ru
Supplementary information:

1 INTRODUCTION

It has always been very important for people to extract bast fibers from the stem of species such as flax and hemp. For extraction bast fibers retting is used. Generally, there are two types of retting: "water retting" and "field-/dew-retting" (Djemiel et al., 2020). In water retting are used anaerobic pectinolytic microorganisms because they are more efficient than aerobic. The most active producers of pectinolytic enzymes (such as polygalactunorases, pectin lyases and pectinesterases) among anaerobic strains are bacteria belonging to the *Clostridium felsineum – Clostridium acetobutylicum* cluster (Tamburini et al., 2004). *Clostridium felsineum DSM 794* was assigned to this cluster by Tamburini et al. (2001). Its close relationship with *Clostridium acetobutylicum* was supported by bootstrap value (100%).

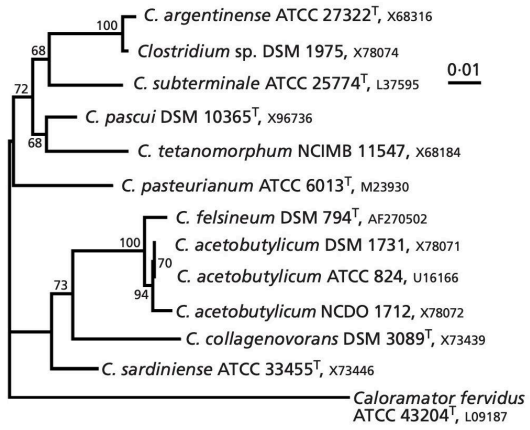

Fig. 1 Unrooted phylogenetic tree based on 16S rDNA comparisons showing the position of *C. felsineum* DSM 794 and representative species of the genus *Clostridium* (Tamburini et al., 2001).

In this mini-review the genome and proteome analysis is presented.

2 METHODS

2.1 Genome analysis

2.1.1 Base pairs' statistics

All tests were carried out on the database from the NCBI database.

The BASH programs were used to calculate the number of nucleotides in different chromosomes of the *Clostridium felsineum DSM 794*. Particularly, the EMBOSS package was used. Files with sequences of definite chromosomes were made with help of the "seqret" program "-ossingle2" option. And after that the number of nucleotides in different chromosomes was counted with help of the "grep" program "-v '^>'" option and "wc" program "-m" option.

Also GC-count was carried out by the "geecee" program from the EMBOSS package.

The number of nucleotides in different chromosomes was counted with my Python program (S5). Sequences from fasta files of the entire genome of *Clostridium felsineum DSM 794*, the main chromosome and the plasmids were used. The number of nucleotides in each of them was counted with help of the list of nucleotides which was passed though the "for" cycle with the".count()" method.

2.1.2 oriC and ter prediction

The oriC and ter were predicted by making a GC-skew plot with help of my Python program (S5). Sequence from the fasta file of the main chromosome was used. The step size 1000 and the window size 100000 were established. The number of G and C nucleotides was counted by the ".count()" method while passing through the sequence with the "for" cycle. Then the GC-skew for each position was counted by the formula: (G - C)/(G + C) (Lu J, Salzberg SL, 2020). Then the cumulative GC-skew was counted by summation of GC-skews of all previous positions to the current position. The maximum and minimum were found by using "max" and "min" functions. Using the "matplotlib" Python library the plot was created.

2.2 Proteome analysis

2.2.1 Protein and RNA analysis

The number of ribosomal proteins was discovered with help of filters in Google Sheets (S1). Firstly, the list with

CDSs (coding sequences) with protein only was made. Secondly, this list was sorted by the presence in the column "name" of "ribosomal protein". But after that not only ribosomal proteins were in this table, ribosomal protein transferases also. Because of that the second filter was made to leave only those names in which there was no "transferase". The ribosomal protein number was counted by the "COUNTUNIQUE" function which was applied to the "ID"column.

The uniqueness of protein was identified by the "IFS" function. If the current cell in the diapason is equal to the next cell "1" is printed in this string in a new column, else "0" is printed. After that the sorting of the table by the new column was made to identify the duplicate values.

The number of transport proteins and hypothetical proteins was counted similarly to the number of ribosomal proteins.

The number of all proteins, all RNAs, rRNAs and tRNAs were calculated by the BASH programs "grep", "cut", "wc". From the feature table of *Clostridium felsineum DSM 794* (S1) in txt format "class" column was taken by "cut" program "-f" option and sorted by "with_protein", "RNA", "tRNA" or "rRNA" with help of the "grep" program. Then the number of strings was calculated by the "wc" program "-l" option.

2.2.2 Protein-coding gene distribution in "+" and "-" strands

The number of genes in different chromosomes in different strands was calculated by sorting the CDS list (S1) in Google Sheets by the columns "chromosome" and "strand" and the "COUNTUNIQUE" function which was applied to the "ID" column. The statistical significance was estimated by "=2*BINOM.DIST(MIN(A3;B3);A3+B3;0,5;TRUE)" formula.

2.2.3 Comparison of the pectin enzymes of *Clostridium felsineum DSM 794* and *Clostridium acetobutylicum DSM 1731*

The comparison was made by sorting with help of Google Sheets *Clostridium felsineum DSM 794* feature table (S1) and *Clostridium acetobutylicum DSM 1731* feature table (S6). These tables were sorted by "pectin" filter in column "name".

3 RESULTS AND DISCUSSION

3.1 Genome analysis

3.1.1 Base pairs' statistics

Clostridium felsineum DSM 794 has one chromosome and two plasmids (p19 and p337). There are 5242194 base pairs in the entire genome: 4885251 base pairs in the main chromosome, 337217 base pairs in the p337 and 19726 base pairs in the p19.

GC-count shows that there are 30% of GC base pairs in the entire genome of *Clostridium felsineum DSM 794*, 30% in the main chromosome, 26.5% in the p19 and 30% in the p337 (S3 and the "geecee" function in BASH). At the same time the closest relative *Clostridium acetobutylicum DSM 1731* has 31% of GC base pairs in the entire genome, 31% in the main chromosome, 31% in the pSMBa and 26% in the pSMBb (S4). It means DNA of *Clostridium acetobutylicum DSM 1731* is more sustainable than DNA of *Clostridium felsineum DSM 794*.

The results of counting the number of nucleotides in different chromosomes of *Clostridium felsimeum DSM 794* are presented in Fig. 2. They confirm the Chargaff's second rule. Only in the p19 significant differences were found (A 41.75% and T 31.8% of the entire genome, G 16.61% and C 9.84% of the entire genome) because the length of this plasmid is small for the reliability of the test.

	the entire genome	
	count	percentage of the entire genome
A	1588687	35.32%
T	1561792	34.72%
G	682466	15.17%
C	665504	14.79%

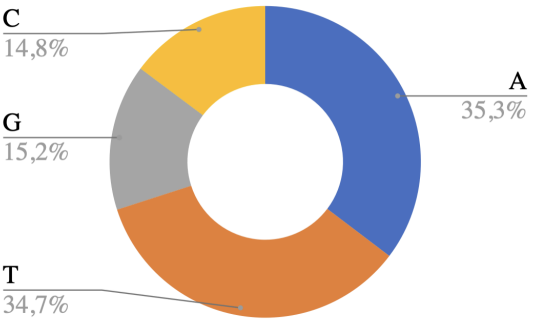

Fig. 2.1 Distribution of nucleotides in the entire genome.

	the main chromosome	
	count	percentage of the entire genome
A	1711502	35.03%
T	1704273	34.89%
G	730855	14.96%
C	738621	15.12%

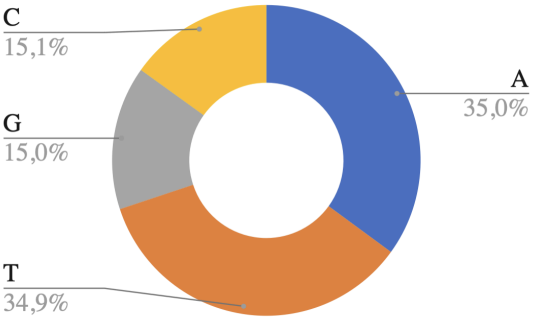

Fig. 2.2 Distribution of nucleotides in the main chromosome.

	p337	
	count	percentage of the entire genome
A	116905	34.67%
T	118638	35.18%
G	50180	14.88%
C	51494	15.27%

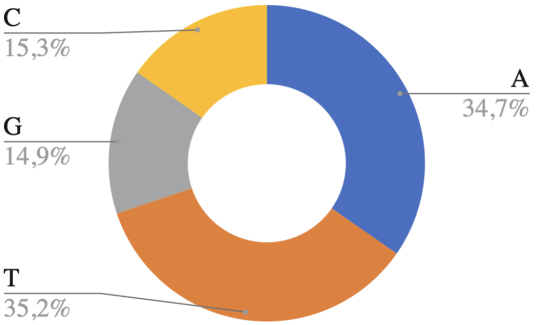

Fig. 2.3 Distribution of nucleotides in the plasmid p337.

	p19	
	count	percentage of the entire genome
A	8235	41.75%
T	6273	31.80%
G	3277	16.61%
C	1941	9.84%

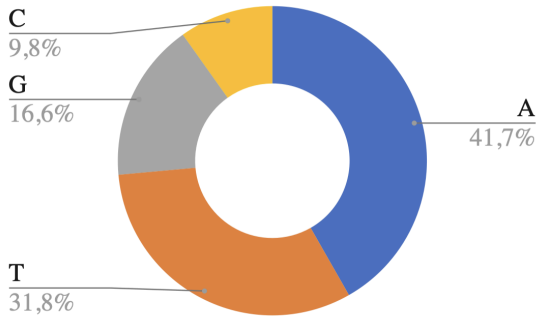

Fig. 2.4 Distribution of nucleotides in the plasmid p19.

3.1.2 oriC and ter prediction

The oriC (start of DNA replication) and ter (termination of DNA replication) were found in the main chromosome (Fig. 3). They were predicted by analyzing the GC-skew plot. The position where GC-skew was the smallest was designated as oriC (approximate position 4785000) and the position where GC-skew was the biggest – ter (approximate position 239500).

3.2 Proteome analysis

3.2.1 Protein and RNA analysis

There are 54 ribosomal proteins in the entire genome. Two of them are with the similar name "30S ribosomal protein S4" and placed in different places in the main chromosome.

Other information about proteins and RNAs and protein length distribution are presented in Fig. 4 and Fig. 6.

As expected almost all RNAs are tRNAs and rRNAs.

The number of hypothetical proteins is huge because the sequence was made recently.

type	count
all proteins	4668
transport proteins	109(2.33% of the entire proteome)
hypothetical proteins	771
all RNAs	118
rRNA	33
tRNA	80

Fig. 4 Another information about proteins and RNAs.

3.2.2 Protein-coding gene distribution in "+" and "-" strands

There are approximately the same number of protein genes on "+" and "-" strands in the main chromosome (91.2% statistical significance) (Fig. 5). In the plasmid p337 also, but statistical significance is low (15.3%) because of the lack of the protein number in the plasmid for credibility. In the plasmid p19 the protein number is even less than in the plasmid p337, so it does not make sense to run the test for statistical significance.

count of genes with protein	chain		statistical significance
	+	-	
the main chromosome	2172	2177	91.2%
p19	18	1	0.0076%
p337	135	165	9.4%

Fig. 5 The protein-coding genes distribution in the "+" and "-" strands in different chromosomes.

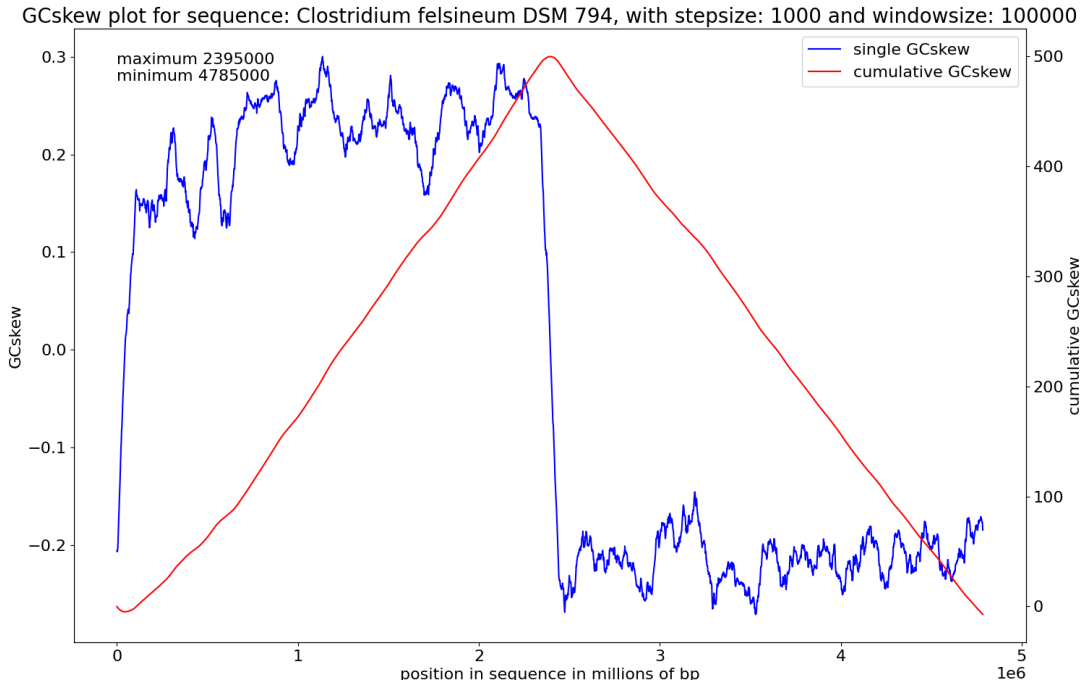

Fig. 3 GC-skew plot of the main chromosome.

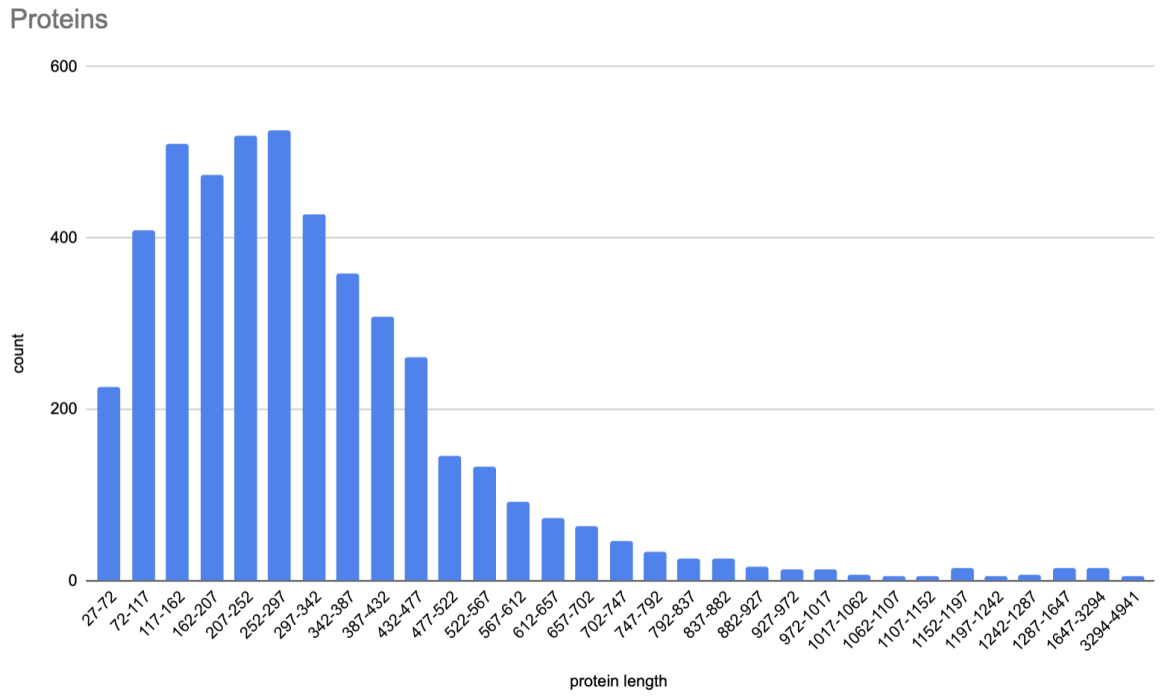

Fig. 6 Protein length distribution.

3.2.3 Comparison of the pectin enzymes of *Clostridium felsineum DSM 794* and *Clostridium acetobutylicum DSM 1731*

It was discovered that the number of the pectin enzymes of *Clostridium felsineum DSM 794* is bigger than *Clostridium acetobutylicum DSM 1731*. *Clostridium felsineum DSM 794* has 6 pectinesterase family proteins in the main chromosome, 1 pectate lyase in the p337 plasmid. *Clostridium acetobutylicum DSM 1731* has 1 pectinesterase family protein in the main chromosome. Both of them have 1 pectate lyase in the main chromosome. From this we can conclude that most likely *Clostridium felsineum DSM 794* is more efficient in retting than *Clostridium acetobutylicum DSM 1731*.

4 CONCLUSION

Clostridium felsineum DSM 794 is a bacterium with 5242194 base pairs in the entire genome. There are 1 chromosome and 2 plasmids. In this research the approximate oriC (4785000 position) and ter (239500 position) in the main chromosome were found. Also the second Chargaff's rule was confirmed.

The distribution of proteins in the chromosomes and protein length was discovered. Also the number of different types of proteins was counted.

The comparison of *Clostridium felsineum DSM 794* and *Clostridium acetobutylicum DSM 1731* was made. Based on my analysis the DNA of *Clostridium acetobutylicum DSM 1731* is more sustainable than *Clostridium felsineum DSM 794* but *Clostridium acetobutylicum DSM 1731* effectiveness of retting is less.

SUPPLEMENTARY MATERIALS

1. *Clostridium felsineum DSM 794* genome feature table.
2. *Clostridium felsineum DSM 794 genome.*
3. *Clostridium felsineum DSM 794* assembly statistics report.
4. *Clostridium acetobutylicum DSM 1731* assembly statistics report.
5. Python programs.
6. *Clostridium acetobutylicum DSM 1731* genome feature table.

REFERENCES

Djemiel, C. et al. (2020). Targeted metagenomics of retting in flax: the beginning of the quest to harness the secret powers of the microbiota. *Frontiers in Genetics*, 11, 1–22.

Tamburini, E. et al. (2001). *Clostridium felsineum* and *Clostridium acetobutylicum* are two distinct species that are phylogenetically closely related. *International Journal of Systematic and Evolutionary Microbiology*, 51, 963-966.

Tamburini, E. et al. (2004). Exploitation of bacterial pectinolytic strains for improvement of hemp water retting. *Euphytica*, 140, 47-54.

Lu J, Salzberg SL (2020) SkewIT: The Skew Index Test for large-scale GC Skew analysis of bacterial genomes. *PLoS Comput Biol*, 16(12): e1008439.